doi:10.1093/nar/gkn279

The Predikin webserver: improved prediction of protein kinase peptide specificity using structural information

Neil F. W. Saunders[1,]* and Bostjan Kobe[2]

[1]School of Molecular and Microbial Sciences and [2]Institute for Molecular Bioscience and Special Research Centre for Functional and Applied Genomics, University of Queensland, Brisbane 4072, Australia

Received January 31, 2008; Revised April 24, 2008; Accepted April 26, 2008

ABSTRACT

The Predikin webserver allows users to predict substrates of protein kinases. The Predikin system is built from three components: a database of protein kinase substrates that links phosphorylation sites with specific protein kinase sequences; a perl module to analyse query protein kinases and a web interface through which users can submit protein kinases for analysis. The Predikin perl module provides methods to (i) locate protein kinase catalytic domains in a sequence, (ii) classify them by type or family, (iii) identify substrate-determining residues, (iv) generate weighted scoring matrices using three different methods, (v) extract putative phosphorylation sites in query substrate sequences and (vi) score phosphorylation sites for a given kinase, using optional filters. The web interface provides user-friendly access to each of these functions and allows users to obtain rapidly a set of predictions that they can export for further analysis. The server is available at http://predikin.biosci.uq.edu.au.

INTRODUCTION

Enzymes of the eukaryotic protein kinase superfamily phosphorylate serine, threonine or tyrosine residues in proteins. Protein kinases and their substrates form complex networks that regulate essentially every eukaryotic cellular process (1). Defects in phosphorylation networks result in numerous disease states, making protein kinases important pharmacological targets. To ensure signalling fidelity, protein kinases act on discrete sets of substrates. Two major factors are responsible for substrate recognition (2): substrate recruitment, encompassing any process that promotes kinase-substrate encounters; and peptide specificity, the preference for particular residues surrounding the phosphorylation site.

We have previously developed a method, named Predikin, to predict the peptide specificity of protein serine–threonine kinases (3). Predikin identifies key conserved *substrate-determining residues* (SDRs) in the protein kinase substrate-binding pocket. The region of the substrate contacted by these residues corresponds to the heptapeptide sequence comprised of positions -3 to $+3$ relative to the phosphorylated residue, so the physicochemical properties of SDRs can be used to predict which heptapeptides are the best substrates for a particular protein kinase. We have recently completely revised and expanded the Predikin codebase and provided access to Predikin via a new webserver. The Predikin webserver is built from three components: (i) Predikin.pm, a Perl module that provides data and methods for the analysis of protein kinase and substrate sequences; (ii) PredikinDB, a database of protein kinases and their substrates and (iii) the website user interface. In this article, we provide a brief description of the methods, capabilities and usage of the Predikin webserver.

METHODS

Protein kinase sequence analysis

Users begin a Predikin analysis by submitting a protein kinase sequence in fasta format. A Perl module, Predikin.pm, provides data and methods for the analysis of both protein kinase and substrate sequences. The protein kinase is analysed using the following methods: (i) assignment of protein kinase type (serine–threonine, CMGC or tyrosine kinase) using a regular expression match based on Prosite patterns (4); (ii) classification by Kinase Sequence Database (KSD) family (5); (iii) classification by PANTHER database family (6) and (iv) identification of the key SDRs. The module makes extensive use of the Bioperl library (7), HMM libraries and the HMMER package (8). SDRs in the query kinase

*To whom correspondence should be addressed. Tel: +61 7 3365 4866; Fax: +61 7 3365 4699; Email: n.saunders@uq.edu.au
Correspondence may also be addressed to Bostjan Kobe. Tel: +61 7 3365 2132; Fax: +61 7 3365 4699; Email: b.kobe@uq.edu.au

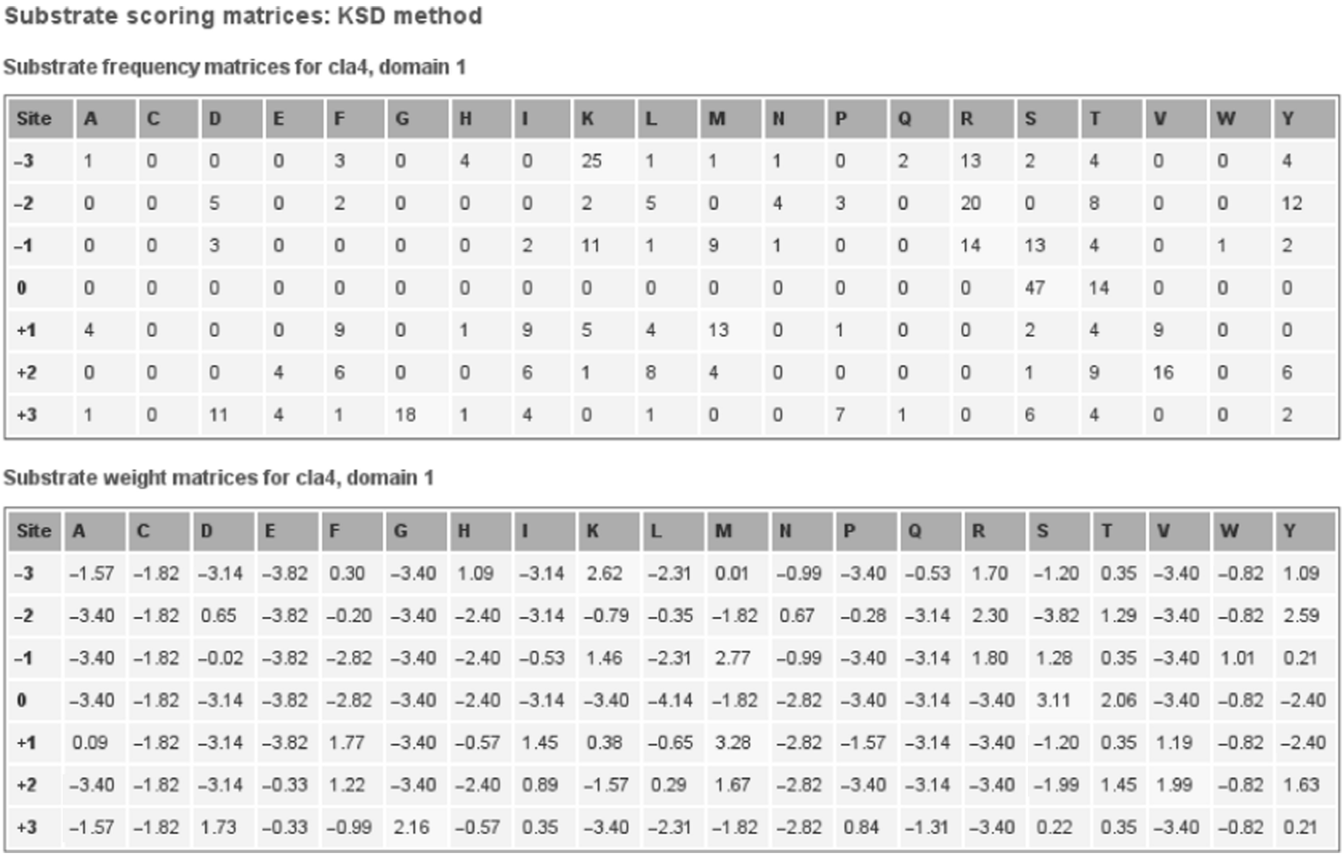
are located using an alignment of the kinase sequence with a HMM profile model of the kinase catalytic domain (S_TKc, accession SM00220) from the SMART database (9). KSD family is assigned using the HMMER tool hmmpfam to compare the kinase sequence with a set of HMMs built from KSD family alignments. PANTHER family is assigned using HMM families and the pantherScore program, both obtained from the PANTHER database website.

Having characterized the catalytic domain of the query kinase, Predikin moves to the next step in the procedure: calculation of kinase-specific weight matrices for substrate prediction.

Kinase-specific weight matrices

The unique feature of Predikin compared with existing methods is that it permits substrate prediction based solely on kinase sequence, as opposed to providing predictions only for a kinase family. This is achieved by querying PredikinDB, the MySQL database backend to the webserver. PredikinDB contains three linked tables that describe protein kinases, their substrates and phosphorylation sites. The data in PredikinDB are derived from the UniProt database using custom parsers written in Perl. PredikinDB is updated automatically at regular intervals using a pipeline of scripts that download UniProt files, parse and generate the database tables. The key feature of PredikinDB is that where possible, phosphorylation sites are linked with the sequence of the kinase acting at the site. This is achieved by parsing the UniProt MOD_RES line for a kinase name (e.g. 'by PKA') and comparing it with a list of gene names for kinases from the same organism as the substrate sequence. PredikinDB currently contains 2335 serine, threonine and tyrosine residues that are annotated as phosphoresidues (with UniProt evidence level 'experimental', 'by similarity', 'probable' or 'potential') in 1116 proteins and that are linked to a specific protein kinase sequence. 11 999 sites are also annotated as experimental by the phospho.ELM database (10), of which 690 are linked to a kinase.

Linking phosphorylation sites with kinase sequences allows the retrieval of phosphorylation sites from the database, where (i) the kinase is known; (ii) the phosphorylation site is annotated with high confidence and (iii) the kinase has similarity in the catalytic domain (as measured by SDRs, KSD or PANTHER family) to those of the query kinase. Users can specify a minimum confidence value for the phosphorylation sites used in scoring matrices (phospho.ELM experimental; UniProt experimental, by similarity, probable or potential) and can also specify that only non-redundant sites be retrieved (homology reduction). The sites are then aligned and used to construct position weight matrices (Figure 1) by comparing the frequency of an amino acid at each position in



Substrate scoring matrices: KSD method

Substrate frequency matrices for cla4, domain 1

Site	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
–3	1	0	0	0	3	0	4	0	25	1	1	1	0	2	13	2	4	0	0	4
–2	0	0	5	0	2	0	0	0	2	5	0	4	3	0	20	0	8	0	0	12
–1	0	0	3	0	0	0	0	2	11	1	9	1	0	0	14	13	4	0	1	2
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	47	14	0	0	0
+1	4	0	0	0	9	0	1	9	5	4	13	0	1	0	0	2	4	9	0	0
+2	0	0	0	4	6	0	0	6	1	8	4	0	0	0	0	1	9	16	0	6
+3	1	0	11	4	1	18	1	4	0	1	0	0	7	1	0	6	4	0	0	2

Substrate weight matrices for cla4, domain 1

Site	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
–3	–1.57	–1.82	–3.14	–3.82	0.30	–3.40	1.09	–3.14	2.62	–2.31	0.01	–0.99	–3.40	–0.53	1.70	–1.20	0.35	–3.40	–0.82	1.09
–2	–3.40	–1.82	0.65	–3.82	–0.20	–3.40	–2.40	–3.14	–0.79	–0.35	–1.82	0.67	–0.28	–3.14	2.30	–3.82	1.29	–3.40	–0.82	2.59
–1	–3.40	–1.82	–0.02	–3.82	–2.82	–3.40	–2.40	–0.53	1.46	–2.31	2.77	–0.99	–3.40	–3.14	1.80	1.28	0.35	–3.40	1.01	0.21
0	–3.40	–1.82	–3.14	–3.82	–2.82	–3.40	–2.40	–3.14	–3.40	–4.14	–1.82	–2.82	–3.40	–3.14	–3.40	3.11	2.06	–3.40	–0.82	–2.40
+1	0.09	–1.82	–3.14	–3.82	1.77	–3.40	–0.57	1.45	0.38	–0.65	3.28	–2.82	–1.57	–3.14	–3.40	–1.20	0.35	1.19	–0.82	–2.40
+2	–3.40	–1.82	–3.14	–0.33	1.22	–3.40	–2.40	0.89	–1.57	0.29	1.67	–2.82	–3.40	–3.14	–3.40	–1.99	1.45	1.99	–0.82	1.63
+3	–1.57	–1.82	1.73	–0.33	–0.99	2.16	–0.57	0.35	–3.40	–2.31	–1.82	–2.82	0.84	–1.31	–3.40	0.22	0.35	–3.40	–0.82	0.21

Figure 1. Frequency (upper) and weight (lower) matrices generated by Predikin to score potential substrates of protein kinase Cla4p from *Saccharomyces cerevisiae*, using the method of classification by KSD family.

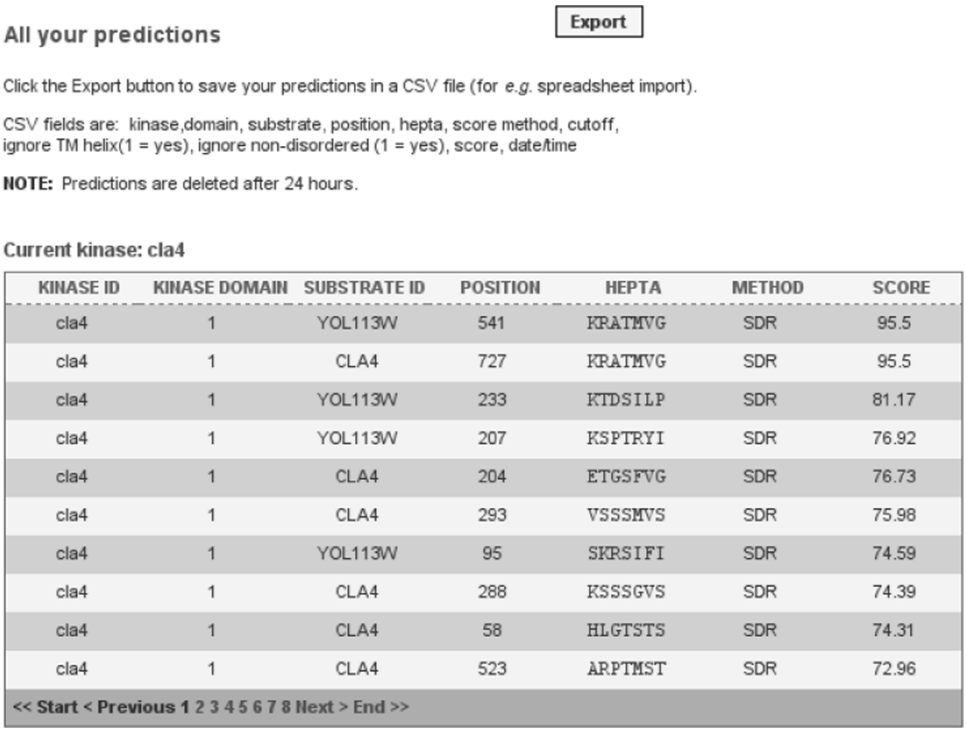
the alignment with the frequency in all substrate sequences for the type (serine–threonine, CMGC or tyrosine kinase) of the query kinase. The matrices can then be used to score potential phosphorylation sites in putative substrates of the kinase.

Substrate prediction

The Predikin.pm Perl module provides methods to score potential phosphorylation sites using the weight matrices generated for the query kinase sequence. The user uploads putative substrates in fasta format, from which all peptides with the sequence XXX[ST]XXX (serine–threonine and CMGC kinases) or XXXYXXX (tyrosine kinases) are extracted. These sites can then be scored using one of the SDR, KSD or PANTHER matrices for the query kinase. A cutoff score below which results are not reported can be specified. In addition, the DisEMBL (11) and TMHMM (12) packages can be employed as filters to discriminate against putative phosphorylation sites on the basis of low intrinsic disorder and location within a transmembrane helix, respectively. Analysis of experimentally validated phosphorylation sites in PredikinDB shows that over 90% are found in a disordered region as predicted by at least one of DisEMBL's three algorithms and <0.1% are located in a TMHMM-predicted helix. The analysis is available as Supplementary data at the website.

The output from Predikin is a table (Figure 2) containing identifiers for the kinase, catalytic domain and substrate, the location of the potential phosphorylated residue, the heptapeptide XXX[STY]XXX and a relative score between 0 and 100 indicating the likelihood of phosphorylation by the kinase. We have performed an evaluation of Predikin scores using kinase-substrate pairs from PredikinDB to determine how well Predikin discriminates known phosphorylation sites of a kinase from unknown sites. The evaluation procedure is provided as Supplementary data at the website. Briefly, sites linked to a kinase were retrieved from PredikinDB and randomly divided into test (10%) and training (90%) sets. All XXX[STY]XXX sites in each test set substrate were scored by generating a scoring matrix for the corresponding kinase, omitting those sites in the training set linked to the same kinase. Known/unknown sites were labelled 1/0, respectively and redundant sites (same peptide, same kinase and so same score) were discarded. The procedure was repeated 100 times to obtain 100 samples of scores and labels for each scoring method/kinase type combination.

Area under receiver operating curve (AROC) values, obtained by plotting true positive (annotated sites) versus false positive (unannotated sites) rates as the score threshold is successively lowered (13) ranged from 0.71 ± 0.98 SD (tyrosine kinases, KSD scores) to 0.93 ± 0.02 SD (CMGC kinases, SDR scores), depending on the Predikin scoring method used and the kinase type. These values indicate that Predikin is effective at distinguishing true sites. Detailed comparison with other methods (GPS, Kinase Phos, NetPhosK, PPSP and Scansite) is beyond the scope of this article; a preliminary AROC analysis indicates that Predikin performs as well or better than other phosphorylation site predictors. However, we emphasize that such comparisons are of limited value, particularly as the other methods can only assign a kinase family to a query



All your predictions

Click the Export button to save your predictions in a CSV file (for *e.g.* spreadsheet import).

CSV fields are: kinase,domain, substrate, position, hepta, score method, cutoff, ignore TM helix(1 = yes), ignore non-disordered (1 = yes), score, date/time

NOTE: Predictions are deleted after 24 hours.

Current kinase: cla4

KINASE ID	KINASE DOMAIN	SUBSTRATE ID	POSITION	HEPTA	METHOD	SCORE
cla4	1	YOL113W	541	KRATMVG	SDR	95.5
cla4	1	CLA4	727	KRATMVG	SDR	95.5
cla4	1	YOL113W	233	KTDSILP	SDR	81.17
cla4	1	YOL113W	207	KSPTRYI	SDR	76.92
cla4	1	CLA4	204	ETGSFVG	SDR	76.73
cla4	1	CLA4	293	VSSSMVS	SDR	75.98
cla4	1	YOL113W	95	SKRSIFI	SDR	74.59
cla4	1	CLA4	288	KSSSGVS	SDR	74.39
cla4	1	CLA4	58	HLGTSTS	SDR	74.31
cla4	1	CLA4	523	ARPTMST	SDR	72.96

Figure 2. A sample prediction generated by the Predikin webserver. Predikin SDR scores for the protein kinase Cla4p from *Saccharomyces cerevisiae* are shown for potential phosphorylation sites in Cla4p and yeast protein YOL113W.

substrate, whereas Predikin predicts substrates based on solely on query kinase sequence.

Web server implementation

The Predikin webserver user interface is built using the open-source Joomla content management system (CMS; http://www.joomla.org). Forms are designed using the Joomla Facile Forms component (http://www.facileforms.biz). The Joomla CMS provides a convenient modular approach to website design, making it easy to add features such as user management, custom forms, documentation and discussion forums. Joomla is written in PHP and so the PECL PHP embedded Perl extension (http://pecl.php.net/perl) is employed to allow communication between the webserver and the Predikin perl module.

The webserver is primarily designed for users interested in a small set of kinases and potential substrates, identified in an experimental screen. However, users can acquire a large set of substrate predictions for a kinase quite rapidly, since all predictions for a session are stored in a temporary database table and can be exported as comma-separated text for easy import to other applications and further analysis. Users with more complex requirements (such as genome-scale prediction of substrates for kinases) may wish to use the standalone Predikin perl module and are encouraged to contact us for more information.

DISCUSSION

The Predikin webserver provides user-friendly access to the improved and enhanced Predikin prediction system. A number of existing tools such as Scansite (14), KinasePhos (15), NetPhosK (16), NetworKIN (17), GPS/GPS2 (18) and PPSP (19) are available to predict protein kinase substrates. The fundamental difference between these tools and Predikin is that analysis using the other tools begins with a substrate sequence, which has to be assigned to a limited number of pre-assigned kinase families. Predikin, on the other hand, uses the kinase sequence to build scoring matrices based on key residues in the kinase catalytic domain that are known from structural analysis to interact with the substrate phosphorylation site. It can therefore make substrate predictions for any protein kinase based on sequence alone, provided that phosphorylation sites of similar kinases are present in the PredikinDB database.

New features and enhancements provided by the revised Predikin code include (i) more reliable determination of SDRs through the use of profile HMM alignments; (ii) filters to prescreen potential phosphorylation sites based on accessibility and disorder; (iii) three methods to generate kinase-specific scoring matrices based on SDRs, KSD or PANTHER family and (iv) use of the PredikinDB database, which is updated continually with new annotated phosphorylation sites and links sites with specific kinase sequences rather than kinase families, so forming the basis for substrate prediction using kinase features.

Predikin provides a range of applications, such as predicting candidate substrates for a protein kinase, candidate protein kinases for a substrate and the assignment of protein kinases to their substrates in large datasets.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Thomas Huber, Ross Brinkworth and Bruce Kemp for helpful discussions. Funding for this work was provided by Australian Research Council Federation Fellowship (FF0561371 to B.K.) and National Health and Medical Research Council Medical Bioinformatics, Genomics and Proteomics Program (395522 to Bruce E. Kemp and B.K.). Funding to pay the Open Access publication charges for this article was provided by the Australian Research Council and National Health and Medical Research Council.

Conflict of interest statement. None declared.

REFERENCES

1. Johnson,S.A. and Hunter,T. (2005) Kinomics: methods for deciphering the kinome. *Nat. Methods*, **2**, 17–25.
2. Kobe,B., Kampmann,T., Forwood,J.K., Listwan,P. and Brinkworth,R.I. (2005) Substrate specificity of protein kinases and computational prediction of substrates. *Biochim. Biophys. Acta*, **1754**, 200–209.
3. Brinkworth,R.I., Breinl,R.A. and Kobe,B. (2003) Structural basis and prediction of substrate specificity in protein serine/threonine kinases. *Proc. Natl Acad. Sci. USA*, **100**, 74–79.
4. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., Castro,E.D., Langendijk-Genevaux,P.S., Pagni,M. and Sigrist,C.J.A. (2006) The PROSITE database. *Nucleic Acids Res.*, **34**, D227–D230.
5. Buzko,O. and Shokat,K. (2002) A kinase sequence database: sequence alignments and family assignment. *Bioinformatics*, **18**, 1274–1275.
6. Mi,H., Lazareva-Ulitsky,B., Loo,R., Kejariwal,A., Vandergriff,J., Rabkin,S., Guo,N., Muruganujan,A., Doremieux,O., Campbell,M.J. *et al.* (2005) The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.*, **33**, D284–D288.
7. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G.R., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
8. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
9. Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J. and Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–D260.
10. Diella,F., Cameron,S., Gemüund,C., Linding,R., Via,A., Kuster,B., Sicheritz-Pontén,T., Blom,N. and Gibson,T.J. (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinform.*, **5**, 79.
11. Linding,R., Jensen,L.J., Diella,F., Bork,P., Gibson,T.J. and Russell,R.B. (2003) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459.
12. Krogh,A., Larsson,B., vonHeijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
13. Sing,T., Sander,O., Beerenwinkel,N. and Lengauer,T. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.

14. Obenauer,J.C., Cantley,L.C. and Yaffe,M.B. (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.*, **31**, 3635–3641.
15. Huang,H.-D., Lee,T.-Y., Tzeng,S.-W. and Horng,J.-T. (2005) KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acid Res.*, **33**, W226–W229.
16. Blom,N., Gammeltoft,S. and Brunak,S. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.*, **294**, 1351–1362.
17. Linding,R., Jensen,L.J., Ostheimer,G.J., vanVugt,M.A.T.M., Jrgensen,C., Miron,I.M., Diella,F., Colwill,K., Taylor,L., Elder,K. *et al.* (2007) Systematic discovery of in vivo phosphorylation networks. *Cell*, **129**, 1415–1426.
18. Xue,Y., Zhou,F., Zhu,M., Ahmed,K., Chen,G. and Yao,X. (2005) GPS: a comprehensive WWW server for phosphorylation sites prediction. *Nucleic Acids Res.*, **33**, W184–W187.
19. Xue,Y., Li,A., Wang,L., Feng,H. and Yao,X. (2006) PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinform.*, **7**, 163.